SEMBED: Semantic Embedding of Egocentric Action Videos

Michael Wray*, Davide Moltisanti*, Walterio Mayol-Cuevas and Dima Damen

Department of Computer Science,
University of Bristol, Bristol, UK
<FirstName>.<LastName>@bristol.ac.uk

Abstract. We present SEMBED, an approach for embedding an egocentric object interaction video in a semantic-visual graph to estimate the probability distribution over its potential semantic labels. When object interactions are annotated using unbounded choice of verbs, we embrace the wealth and ambiguity of these labels by capturing the semantic relationships as well as the visual similarities over motion and appearance features. We show how SEMBED can interpret a challenging dataset of 1225 freely annotated egocentric videos, outperforming SVM classification by more than 5%.

Keywords: Egocentric Action Recognition, Semantic Ambiguity, Semantic Embedding

1 Introduction

An egocentric camera captures rich and varied information of how the wearer interacts with their environment. The challenge for the visual understanding of this information is currently significant and not only incited by the enormous variety of such interactions but also by limitations in the available visual descriptors, e.g. those rooted in motion or appearance. Supervised learning from labelled examples is used to alleviate some of these ambiguities. Egocentric datasets [12, 10, 34, 6] and interaction recognition methods [10, 28, 9, 23] differ in the features used and classification techniques adopted, yet they all assume a semantically distinct set of *pre-selected* verbs or verb-noun combinations for supervision. When free annotations are available - unbounded choice of verbs or verb-nouns - from audio scripts [1] or textual annotations [6], a single label is selected to represent each interaction using a majority vote. Less frequent annotations are treated as outliers, though they typically represent a meaningful and correct annotation. For example, lifting an object from a workspace could be described as *pick-up*, *lift*, *take* or *grab*; all valid labels. Note that assuming multiple *valid* labels is different from the problem of Ambiguous Label Learning, [3, 14], where the aim is to find a single valid label from a mixed set of related and unrelated labels.

* Both authors contributed equally to this work

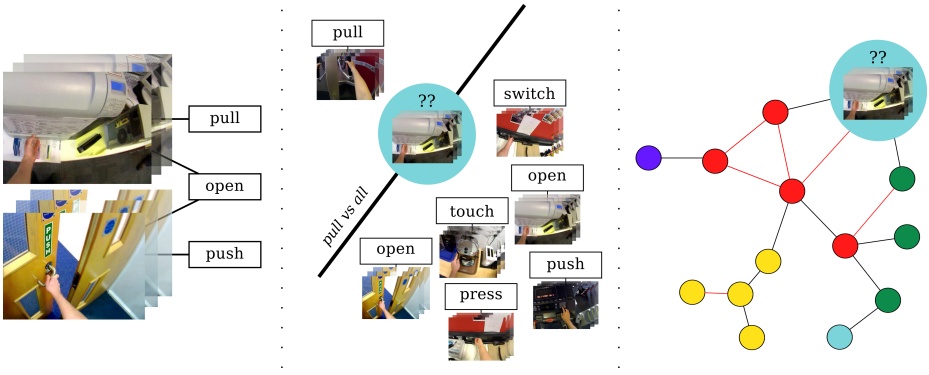

Fig. 1. Given a dataset of free annotations, with potentially ambiguous semantic labelling (left), we propose to deviate from the one-vs-all classical approach (middle) and instead build a graph that encapsulates semantic relationships and visual similarities in the training set (right). Recognition then amounts to embedding an unlabelled video (denoted by '??') into the graph and estimating the probability distribution over potential labels.

Egocentric video offers a unique insight into object interactions in particular. The camera is ideally positioned to capture objects being used and, equally interesting, the different ways in which the same object is used. One interaction (e.g. *open*) applies to a wide variety of objects, and each video can be labelled by multiple valid labels (e.g. *open door* vs *push door*). In this context, recognition cannot be simplified as a one-vs-all classification task. Capturing the semantic relationships between annotations and the visual ambiguities between accompanying video segments can better represent the space of possible interactions. Figure 1 shows a graphical abstract of our work.

Given a dataset of egocentric object interactions with free annotations, we contribute four diversions from previous attempts: (i) We treat all free annotations as valid, correct labellings, (ii) A graph that combines semantic relationships with visual similarities is built, inspired by previous work on object class categories in single images [8] (Sec. 3.1), (iii) A test video is embedded into the previously learnt semantic-visual graph and the probability distribution over its possible annotations is estimated (Sec. 3.2) and (iv) When verb meanings are available, we discover semantic relationships between annotations using WordNet (Sec. 3.3).

We test semantic embedding (SEMBED) on three public egocentric datasets [6, 34, 9]. We show that as the number of verb annotations and their semantic ambiguities increase, SEMBED outperforms classification approaches. We also show that incorporating higher level semantic relationships, such as the hyponymy relationship, improves the results. Note that while we focus on *egocentric object interaction recognition* as a rich domain of semantic and visual ambiguities, some of the arguments can apply to action recognition in general.

2 Embedding Object Interactions - Prior Work

To the best of our knowledge, embedding for egocentric action recognition has not been attempted previously. We first review works on recognising egocentric object interactions, then review works which incorporate semantic knowledge for recognition tasks.

Egocentric Object Interaction Recognition: Egocentric action recognition works range from self-motion [17] (e.g. walk, cycle) to high-level activities (e.g. [34, 18, 20, 2, 35]). On the task of object interaction recognition, approaches vary in whether they use hand-centred features [15, 19], object-specific features [10, 6, 23, 29] or a combination [12, 21]. Ishihara *et al* [15] use dense trajectories in addition to global hand shape features and apply a linear SVM to determine the action class. Kumar *et al* [19] sample and describe superpixel regions around the hand. Their method allows for hand detectors to be trained spontaneously with the user performing the action.

Object-specific features are better suited for recognising verb-noun actions (e.g. *pick-cup vs pick-plate*) rather than a general *picking* action. In Damen *et al* [6], spatio-temporal interest points have been used to discover object interactions in an unsupervised manner. The works of Fathi *et al* [10, 9, 21, 11] have tested features including gaze, colour, texture and shape for verb-noun action classification. Of these, [10] specifically discusses the change in the object state as a useful feature to recognise object interactions. Though attempting video summarisation primarily, Ghosh *et al* [12] introduces a collection of features that could be used to classify object-interactions such as distance from the hand, saliency, objectness represented using a spatio-temporal pyramid to detect change. These features were proven useful for segmenting object-interactions from a lengthy video, but have not been tested for action classification *per se*. On several publicly available datasets, Li *et al* [21] compare motion, object, head motion and gaze information along with a linear SVM for object interaction classification. Their results prove that Improved Dense Trajectories (IDT) proposed by [37] outperform other motion features.

With the emergence of highly-discriminative appearance-based features, pre-trained Convolutional Neural Networks (CNN) on ImageNet have also been tested. In [25], CNN is evaluated for distinguishing manipulation from non-manipulation actions on an RGB-D egocentric dataset. Ryoo *et al* [30] combine CNN with IDT along with novel time series pooling for dog-centric manipulation and non-manipulation actions. More recently, fine-tuned multi-stream CNN approaches have achieved state of the art results on egocentric datasets [22, 33], though are tuned on each dataset independently.

Based on [21, 30] conclusions, in this work we report results on IDT as a state-of-the-art motion feature and pre-trained CNN features a state-of-the-art appearance feature. Testing tuned CNNs is left for future work.

Semantic Embedding for Object and Action Recognition: Using linguistic semantic knowledge for Computer Vision tasks, including action recog-

nition, has been fuelled by the accessibility of text or audio descriptions from online sources.

One such dataset which made this possible was gathered from YouTube videos [4] with free annotations. The dataset includes a variety of real-world scenarios, though not limited to egocentric or object-interactions. For each video, multiple annotators were asked to describe the video. Both [26, 13] use this dataset for action recognition. In Motwani and Mooney [26], the most frequently annotated verb for each video is used, and verbs are grouped into classes using semantic similarity measures, extracted from the WordNet hierarchy as well as information corpuses. Videos are described by HoG and HoF features around spatio-temporal interest points. Guadarrama *et al* [13] find subject, object and verb triplets in an attempt to automatically annotate the action. They create a separate semantic hierarchy for each, formulated by co-occurrences of words within the free annotations and use Spearman's rank to find the distances between clusters. Semantic links are used to generate specific, rather than general, annotations and a classifier is trained for each leaf node within the hierarchies. Their method allows zero-shot action annotation by trading-off specificity and semantic similarity. While combining semantics, both works use majority voting to limit the description per class to a single verb.

Another recent YouTube dataset was collected of users performing tasks while narrating their actions [1]. Labels are extracted from audio descriptions using automatic speech recognition. Verb labels are then used to align videos using a WordNet similarity measure as well as visual similarity (HoF and CNN) to find the sequence of actions in a task.

Semantics have also been used for object recognition with images. Jin *et al*[16] use WordNet to remove noisy labels from images which have multiple labels. Similarly, Ordonez *et al* [27] use WordNet to find the most frequently-used object labels amongst multiple annotations. We build our work on Fang and Torresani [8], where images are embedded in a semantic-visual graph. In [8], images are clustered depending on the semantic relationships between the labels and edges of the graph are weighted with the visual similarity. They use ImageNet as the database for training, and benefit from the fact that images within ImageNet are organised according to the WordNet hierarchy. We differ from [8] in how we add visual links to the semantic graphs as will be explained next.

3 Semantic Embedding of Egocentric Action Videos

We next, in Sec. 3.1, explain how we build a semantic-visual graph (SVG) that encodes label and visual ambiguities in the training set. In Sec. 3.2, we detail how videos with an unknown class are embedded in SVG, and how the probability distribution over their annotations is estimated. Finally, in Sec. 3.3 we explore further semantic relationships when verb meanings are annotated.

3.1 Learning the Semantic-Visual Graph

The Semantic-Visual Graph (SVG) is a representation of the training videos, with three sources of information encoded. First, videos that are semantically linked, e.g. have the same label, are linked in SVG. Second, nodes that are visually similar, yet semantically distinct, should also be linked as these indicate visual ambiguities. Third, edge weights correspond to the normalised visual similarity, over neighbouring nodes, using a visual descriptor and a defined distance measure. In this section we explain how SVG_u, an undirected graph, is constructed, then normalised to achieve the directed graph SVG.

SVG_u is an undirected graph, where one node $x_i \in \text{SVG}_u$ corresponds to one training video. Assume $\text{AX}(x_i, x_j)$ is a binary function that checks whether two video labels are semantically related. Initially, $\text{AX}(x_i, x_j)$ is *true* when both videos are annotated by the exact same verb. This assumption is revisited in Sec. 3.3. Edges in SVG_u are created between nodes with a semantic relationship:

$$x_i \frown x_j \in \text{SVG}_u \iff AX(x_i, x_j) = true \tag{1}$$

The undirected edge $x_i \frown x_j \in \text{SVG}_u$ is assigned a weight $w_{x_i \frown x_j} = D_v(x_i, x_j)$ where D_v is a distance measure defined over the visual descriptor chosen. Assume $rank(D_v(x_i, x_j))$ is a function that returns the relative position of the distance measure amongst all the remaining pairs of videos such that,

$$rank(D_v(x_i, x_j)) = n \iff D_v(x_i, x_j) = min_n(D_v(x_k, x_l)) \quad \forall x_k, x_l \in \text{SVG}_u \quad and \quad AX(x_k, x_l) \neq true \tag{2}$$

and min_n is the n^{th} minimum element in the list. In addition, assume $rank_i(D_v(x_i, x_j))$ is a function that returns the relative position of $D_v(x_i, x_j)$ amongst all nodes not connected to x_i such that,

$$rank_i(D_v(x_i, x_j)) = n \iff D_v(x_i, x_j) = min_n(D_v(x_i, x_l)) \quad \forall x_l \in \text{SVG}_u \quad and \quad AX(x_i, x_l) \neq true \tag{3}$$

Further links are added to SVG_u to encode visual ambiguities such that,

$$x_i \frown x_j \in \text{SVG}_u \iff rank(D_v(x_i, x_j)) \leq m \quad or \quad rank_i(D_v(x_i, x_j)) = 1 \tag{4}$$

where m is the number of visual connections in SVG_u that correspond to the top m visually similar and semantically dissimilar nodes in SVG_u. We differ from [8] in that we ensure each node is connected to its top visually similar but semantically distinct node.

The undirected graph SVG_u is then converted to a directed graph by replacing each edge with two directed edges.

$$x_i \frown x_j \in \text{SVG}_u \Rightarrow \{x_i \rightarrow x_j, x_j \rightarrow x_i\} \in \text{SVG} \tag{5}$$

The weights of directed edges are initially the same as the weights for their undirected counterparts however they are normalised to define the probability of traversing from video x_i to x_j,

$$P(x_i \rightarrow x_j) = \frac{1/w_{x_i \rightarrow x_j}}{\sum_k 1/w_{x_i \rightarrow x_k}} \qquad \forall x_i \rightarrow x_k \in \text{SVG} \tag{6}$$

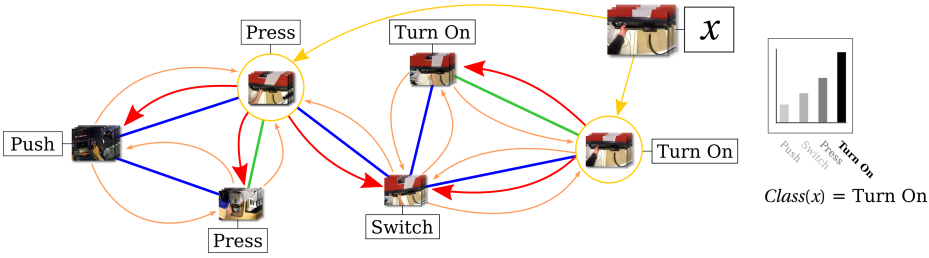

Fig. 2. The Semantic-Visual Graph (SVG) is built for training data, with semantic links (green) and visual links (blue) between videos. Given a test video x, two nearest neighbours are found (yellow) and a Markov Walk of 2 steps (step1-red and step2-orange) finds the probability distribution over potential labellings. Ref. supplementary material for animation.

The reciprocal of the weights is taken so that the most visually similar path will have the highest probability.

3.2 Embedding in Semantic-Visual Graph

Given a test video, x, we first embed the video into SVG then use the Markov Walk (MW) method from [8] to determine $Class\,(x)$. To embed x, we begin by finding the set $\mathcal{R}$ which contains the z closest neighbours to x based on visual distance, such that

$$\mathcal{R} = \{x_i \in \text{SVG} \mid rank(D_v(x, x_j)) \leq z\} \tag{7}$$

We embed x into SVG by adding directed edges connecting x to nodes in $\mathcal{R}$: $x \rightarrow x_i \quad \forall x_i \in \mathcal{R}$ with normalised weights $P(x \rightarrow x_i)$. Following the embedding, MW attempts to traverse the nodes in the directed graph to estimate the probability of $Class(x)$. Given the Markovian assumption and a predefined number of steps t, we calculate the probability distribution of reaching a node x_i as follows

$$P(x_{i+t} \mid x) = \prod_{x_i \in R} \left(P(x \rightarrow x_i) \prod_{j=1}^{t} P(x_{i+j-1} \rightarrow x_{i+j}) \right) \tag{8}$$

To perform MW efficiently, we construct the vector q such that

$$q(i) = \begin{cases} P(x \rightarrow x_i) & x_i \in \mathcal{R} \\ 0 & \text{otherwise} \end{cases} \tag{9}$$

We also construct a matrix A such that $A(i, j) = P(x_i \rightarrow x_j)$ (Eq. 6), note that this matrix is asymmetrical as nodes have a different set of neighbours in SVG. Accordingly, $P(x_{i+t} \mid x) = q^T A^t$ where q^T is the transpose of q and t is the

number of steps in MW. We can then accumulate $P(Class(x))$ for every unique annotation $ax \in AX$ as follows

$$P(Class(x) = ax) = \sum_{AX(x_{i+t}, ax)=true} P(x_{i+t} \mid x) \tag{10}$$

We then select $\arg\max_{Class(x)} P(Class(x))$ as the semantic label of x. Figure 2 shows an example of SVG and video embedding. In the figure, given two nearest neighbours $z = 2$ and two steps in MW $t = 2$, the probability distribution over possible labellings is calculated.

3.3 Semantic Relationships: Synsets and Hyponyms

In Sec. 3.1, videos are considered semantically linked only when the annotated verbs are the same. SVG then enables handling ambiguities via incorporating visual similarity links in the graph. However, further semantic relationships, such as synonymy and hyponymy relationships, can be exploited between annotations. In linguistics, two words are *synonyms* if they have the same meaning, and the set of all synonyms is a *synset*. Moreover, two words are described as a *hyponym* and a *hypernym* respectively if the first is a more specific instance of the second. The terms originate from the Greek word *hypó* and *hypér* - *under* and *over*.

Synonymy and hyponymy relationships are encoded in lexical databases. WordNet (v3.1, 2012) is a commonly-used lexical database that is based on six semantic relations [24]. In the WordNet verb hierarchy, verbs are first separated into their various meanings by the notation $\langle word \rangle.v.\langle s \rangle$ where $s \geq 1$ is the number of disjoint meanings. The meanings are then arranged in hierarchies that encapsulate semantic relationships. To benefit from such hierarchies, verbs should be annotated with their meanings. We annotate [6] using verb meanings, and Fig. 3 shows how such annotations of the same action can be synonyms and hyponyms, as annotators chose different or more specific action descriptions.

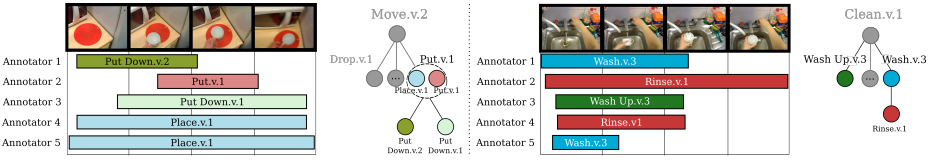

Fig. 3. Five free annotations for two sequences from the BEOID dataset [6], and the respective semantic relationships between the annotations from WordNet [24]. In the hierarchy, each parent-child relationship represents a hypernym-hyponym pair. The dotted circle encapsulates a synonymy relationship. The start and end times of the actions are also shown. For placing a cup on a mat (left), synonyms *put.v.*1 and *place.v.*1 were chosen by annotators. *put_down.v.*1, a hyponym of *put.v.*1 was also used. For washing a cup (right), the verbs *wash.v.*3, *wash up.v.*3 and *rinse.v.*1 were chosen. *rinse.v.*1 is a hyponym of *wash.v.*3.

Name	Users	Seq.	OI Seg.s	Used OI Seg.s	Semantic Verbs
CMU [7]	5	35	516	406	12 (33.8, 30.5)
GTEA+ [9]	13	30	3371	1000⋆	25 (40.0, 75.5)
BEOID [6]	3-5	58	1488	1225	75 (16.3 34.2)

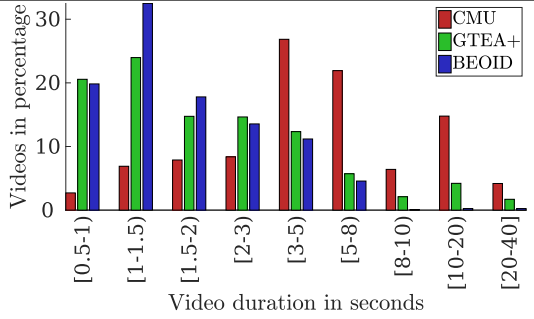

Fig. 4. Dataset details (top) and video length distributions (bottom). Number of users, segments, Object-Interaction (OI) segments and used segments in the results (length < 40s) are detailed. We report the number of annotated verbs along with μ and σ for the number of segments per verb. *: Due to the size of GTEA+ we sampled 1000 videos randomly. Ref. supplementary material for frequencies of verb annotations per dataset.

Given annotated meanings, we define the term action synsets (AS) to indicate that annotations are linked by a synonymy relationship solely, and the term action hyponym (AH) to indicate that annotations are linked by both the synonymy or the hyponymy relationships. For comparison, we define the term action meaning (AM) where annotations are linked only when the annotation matches exactly. We use the general term AX where $AX \in \{\text{AM, AS, AH}\}$ is one of the the possible types of semantic relationships tested.

4 Datasets, Experiments and Results

We selected three publicly available datasets that primarily focus on object interactions from egocentric videos [7, 9, 6] (Figure 4).

Verb annotations: We exploited the annotations provided by the authors to split the CMU and GTEA+ sequences into object-interaction segments. For CMU, object-interaction annotations are only provided for the activity of *making brownies*. Annotators chose from 12 disjoint verbs to ground-truth segments. In GTEA+ annotators chose from verb-noun pairing to ground-truth, e.g. *cut_cucumber* versus *divide_bun* and similarly *squeeze_ketchup* versus *compress_bun*. When removing the nouns, verbs could be used interchangeably but free annotations were not available to annotators.

While BEOID contains a variety of activities and locations, ranging from a desktop to operating a gym machine, it does not provide action-level annota-

Table 1. As the number of verbs increases from 12 to 75, the best performance changes from SVM to SEMBED. Results are obtained with $\gamma_{fv} = 10$ and $\gamma_{bow} = 256$, k ={3,5,5}, $m = 240$, z={2,6,4}, t={20,20,8} for CNN and z={4,5,14}, t={4,20,10} for IDT. For completion, state-of-the-art results on verb-noun classes are reported under 'Other Works' thus are not directly comparable to our verb only results.

FEATURES	CNN						IDT							
ENCODING	FV			BOW			FV			BOW				
METHOD	SVM	K-NN	SEMBED	SVM	K-NN	SEMBED	SVM	K-NN	SEMBED	SVM	K-NN	SEMBED	Verbs	Other Works
CMU [7]	58.6	46.6	46.3	55.9	43.3	52.0	**69.4**	58.1	57.4	55.9	57.6	61.6	12	48.6 [34], 73.4 [36]
GTEA+[9]	15.6	30.0	31.0	25.1	33.5	33.6	**43.6**	**43.4**	42.1	27.8	34.5	40.3	25	60.5 [21], 65.1 [22]
BEOID [6]	20.9	34.4	37.5	15.2	19.1	19.6	38.7	36.0	37.4	34.8	39.6	**45.0**	75	-

tions so we annotated BEOID using free annotations[1], allowing annotators to split video sequences into object-interaction segments in addition to choosing the verb. We recruited 20 native English speakers. These annotators were given a free textbox to label each segment with the verb that best described the seen interaction *in their opinion*. Once a verb has been chosen, the annotators were given the set of potential meanings extracted from WordNet for the chosen verb. Again, they were asked to select the meaning that, *in their opinion*, best suited the segment. Multiple annotators (8-10) were asked to label each task to intentionally introduce variability in the choice of verbs and start-end times of object interaction segments.

Motion and Appearance Features: We test two state-of-the-art feature descriptors to represent both the motion and the appearance of the videos. These are the Improved Dense Trajectories (IDT) [38] and Overfeat Convolutional Neural Networks pre-trained for ImageNet classes (CNN) [32]. For CNN features, we take every 5th frame from 30fps video, starting always from the first frame, and rescale to 320x240 pixels.

Encodings: We test two encodings, using Bag of Words (BoW) [5] and Fisher Vectors (FV) [31] with Euclidean distance. For IDT, when creating the BoW and FV representations, we use a 25% random sample from every video to model the Gaussians for efficiency. We vary the number of Gaussians (γ_{fv}) and the size of the codebook (γ_{bow}) in reported results.

Classification: In all results, leave-one-person-out cross validation has been used. Namely, when testing a video containing one person performing an action, all other videos captured from the same person are excluded from the training set. For SVM results, as the tested datasets contain an imbalance in the distribution of instances per class, we weight the classes by the term $w(c) = 1/prior(c)^{\lambda}$ where $\lambda \in [0, 1]$ is the exponent that best fits the distribution of segments per verb for a given dataset (ref supplementary material).

Results on annotated verbs: Table 1 compares the three datasets for every ⟨features, encoding, classifier⟩ combination. The following conclusions can be made: (i) for all datasets, motion features (IDT) outperform appearance features (CNN) when classifying verbs without considering the object used. (ii) for CMU and GTEA+, we produce comparable results to published results us-

[1] Annotations can be found at: `http://www.cs.bris.ac.uk/~damen/BEOID/`

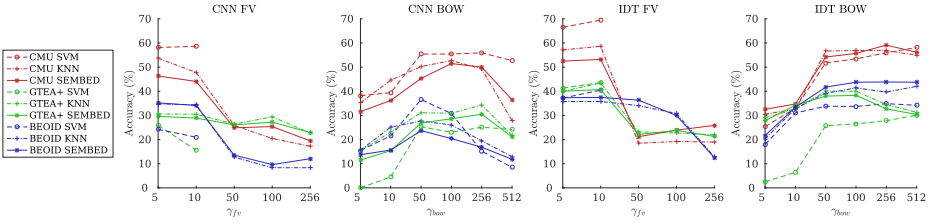

Fig. 5. Results as γ_{fv} and γbow vary for CMU, GTEA+, BEOID. Results were shown with $k = 5$, $m = 240$, $z = 10$, $t = 10$. Similar performance is seen for other parameters.

ing motion information on the same datasets. These are reported under 'Other Works' but are not directly comparable as published works tend to report on verb-noun classes. (iii) For the three datasets with varying number of verbs, as the number of verbs increases (12 → 75) with an increase in semantic ambiguity, SEMBED outperforms standard classifiers (SVM and K-NN). While the table shows the best results for encoding, Fig. 5 reports comparative results as γ is changed - $\gamma_{fv} = 10$ generally led to higher accuracies on all datasets, compared to $\gamma_{bow} = 256$.

We test the sensitivity of SEMBED to its key parameters z and t and report results in Fig. 6 showing the accuracy over various features for BEOID and across the three datasets for IDT-BOW (Ref. supplementary material for all combinations). As noted, z and t behave differently for the various appearance and motion descriptors as well as for different encodings. Generally, SEMBED is more sensitive to the choice of z than t. This is because the Markovian Walk (MW) is unable to represent the probability distribution over labels unless the starting positions are representative of the visual ambiguity. Figure 6 also shows that MW isn't too helpful for CMU (as t increases, accuracy decreases) because it has visually distinctive verb classes. On all datasets, SEMBED is resilient to changing m values; the results are comparable on $180 \leq m \leq 400$.

Results on annotated verbs and meanings: As mentioned earlier, we also annotate BEOID with verb-meaning ground-truth. This resulted in 108 $\langle word \rangle.v.\langle s \rangle$ annotations for the 1225 segments in the dataset. Note the increase in the number of classes from 75 when using verbs only to 108 when using verb-meaning ground-truth. This increase is due to two reasons - one *helpful*, another *problematic*. For example, it is *helpful* when annotators choose between *hold.v*.1: *"keep in a certain state, position"* and *hold.v*.2: *"hold in one's hand"*. Annota-

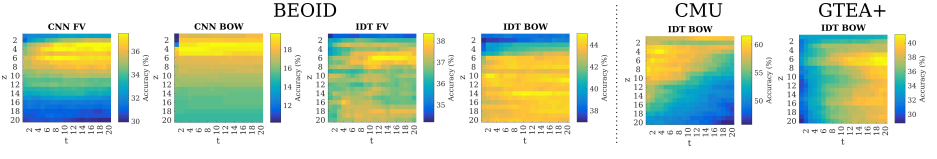

Fig. 6. Evaluation of SEMBED sensitivity to z and t parameters with $m = 240$.

Table 2. As synonymy (AS) and then hyponymy (AH) semantic relationships are incorporated, accuracy increases for all features on the BEOID dataset. $\gamma_{fv} = 10$, $\gamma_{bow} = 256$, $m = 240$, {AM,AS,AH}: z_{CNN}={3,3,2}, t_{CNN}={20,20,14}, z_{IDT}={6,10,13}, t_{IDT}={20,20,2}.

FEATURES	CNN						IDT						
ENCODING	FV			BOW			FV			BOW			
METHOD	SVM	K-NN	SEMBED	SVM	K-NN	SEMBED	SVM	K-NN	SEMBED	SVM	K-NN	SEMBED	Classes
AM	13.2	24.6	26.2	12.1	7.8	11.7	25.9	28.5	32.2	26.1	31.6	**38.2**	108
AS	17.9	25.6	27.1	12.7	8.1	12.7	29.8	30.4	33.5	29.6	33.6	**40.6**	102
AH	18.1	25.0	26.9	12.2	7.4	16.3	36.2	33.1	34.5	29.1	35.2	**41.9**	84

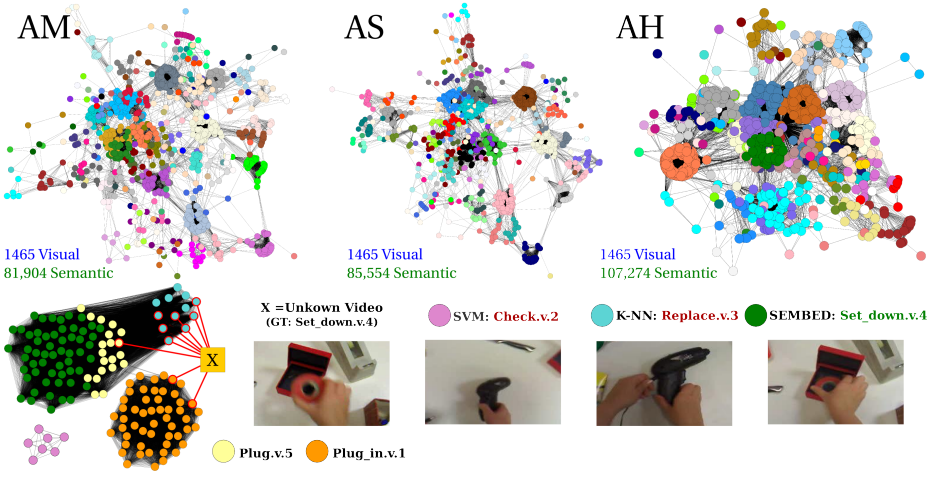

Fig. 7. SVG for three semantic levels on BEOID (top). Example using AH (bottom), SVM and K-NN produce incorrect results. The Markov walk of SEMBED allows the video to be correctly classified.

tors would then use the first for when a button is pressed and the second for when an object is grasped. However, frequently, WordNet meanings can appear ambiguous resulting in *problematic* cases, especially in the context of egocentric actions. An example of this is the action of turning a tap on so water would flow. Annotators used *turn.v.*1: *"change orientation or direction"* and *turn.v.*4: *"cause to move around or rotate"* interchangeably. In WordNet though, *turn.v.*1 and *turn.v.*4 are not semantically related, introducing unwanted ambiguity affecting the ground-truth labels. While we accept that WordNet may not be the best method to incorporate meaning, we report results as semantic links are incorporated.

We test the three types of semantic relationships AX = {AM,AS,AH}. Histograms of all classes for the various semantic relationships are included in the supplementary material. Table 2 shows that embedding consistently improved performance as synsets and then hypernyms are grouped. Results also demonstrate the advantages of introducing semantic links between videos. Additionally, IDT continues to outperform CNN. Figure 7 shows one example of SEMBED in

action when using meanings and AH semantic links[2]. It should be noted that the best performance of SEMBED on meanings is inferior to using verbs only. This is due to the difficulty in assigning meanings to verbs as previously noted. Approaches to address meaning ambiguities are left for future work.

5 Conclusion and Future Directions

The paper proposes embedding an egocentric action video in a semantic-visual graph to estimate the probability distribution over potentially ambiguous labels. SEMBED profits from semantic knowledge to capture interchangeable labels for the same action, along with similarities in visual descriptors.

While showing clear potential, outperforming classification approaches on a challenging dataset, results merely evaluate the arg max label when compared to ground-truth. Further analysis of the probability distribution will be targeted next. Other approaches to identify semantically related object-interaction labels from, for example, other lexical sources, overlapping annotations or object labels will also be attempted. SEMBED's ability to scale to other object interactions and more discriminative visual descriptors will also be tested.

References

1. Alayrac, J., Bojanowski, P., Agrawal, N., Laptev, I., Sivic, J., Lacoste-Julien, S.: Unsupervised learning from narrated instruction videos. In: CVPR (2016)
2. Bleser, G., Damen, D., Behera, A., Hendeby, G., Mura, K., Miezal, M., Gee, A., Petersen, N., Macaes, G., Domingues, H., Gorecky, D., Almeida, L., Mayol-Cuevas, W., Calways, A., Cohen, A., Hogg, D., Stricker, D.: Cognitive learning, monitoring and assistance of industrial workflows using egocentric sensor networks. PLOS ONE (2015)
3. Chen, C.H., Patel, V.M., Chellappa, R.: Matrix completion for resolving label ambiguity. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2015)
4. Chen, D., Dolan, W.: Collecting highly parallel data for paraphrase evaluation. In: Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (2011)
5. Csurka, G., Dance, C., Fan, L., Willamowski, J., Bray, C.: Visual categorization with bags of keypoints. In: Workshop on statistical learning in computer vision, ECCV (2004)
6. Damen, D., Leelasawassuk, T., Haines, O., Calway, A., Mayol-Cuevas, W.: You-do, I-learn: Discovering task relevant objects and their modes of interaction from multi-user egocentric video. In: BMVC (2014)
7. De La Torre, F., Hodgins, J., Bargteil, A., Martin, X., Macey, J., Collado, A., Beltran, P.: Guide to the Carnegie Mellon University Multimodal Activity (CMU-MMAC) database. Robotics Institute (2008)
8. Fang, C., Torresani, L.: Measuring image distances via embedding in a semantic manifold. In: ECCV (2012)

[2] Video with results available at: `http://youtu.be/6bDDTIJUuic`

9. Fathi, A., Li, Y., Rehg, J.: Learning to recognize daily actions using gaze. In: ECCV (2012)
10. Fathi, A., Rehg, J.: Modeling actions through state changes. In: CVPR (2013)
11. Fathi, A., Ren, X., Rehg, J.: Learning to recognize objects in egocentric activities. In: CVPR (2011)
12. Ghosh, J., Lee, Y.J., Grauman, K.: Discovering important people and objects for egocentric video summarization. In: CVPR (2012)
13. Guadarrama, S., Krishnamoorthy, N., Malkarnenkar, G., Venugopalan, S., Mooney, R., Darrell, T., Saenko, K.: Youtube2text: Recognizing and describing arbitrary activities using semantic hierarchies and zero-shot recognition. In: ICCV (2013)
14. Hüllermeier, E., Beringer, J.: Learning from ambiguously labeled examples. Intelligent Data Analysis pp. 419–439 (2006)
15. Ishihara, T., Kitani, K., Ma, W., Takagi, H., Asahawa, C.: Recognizing hand-object interactions in wearable camera videos. In: ICIP (2015)
16. Jin, Y., Khan, L., Wang, L., Awad, M.: Image annotations by combining multiple evidence & Wordnet. In: ACM international conference on Multimedia (2005)
17. Kitani, K., Okabe, T., Sato, Y., Sugimoto, A.: Fast unsupervised ego-action learning for first-person sports videos. In: CVPR (2011)
18. Kuehne, H., Serre, T.: Towards a generative approach to activity recognition and segmentation. AXiv preprint ArXiv:1509.01947 (2015)
19. Kumar, J., Li, Q., Kyal, S., Bernal, E., Bala, R.: On-the-fly hand detection training with application in egocentric action recognition. In: CVPRW (2015)
20. Lade, P., Krishnan, N., Panchanathan, S.: Task prediction in cooking activities using hierarchical state space markov chain and object based task grouping. In: ISM (2010)
21. Li, Y., Ye, Z., Rehg, J.: Delving into egocentric actions. In: CVPR (2015)
22. Ma, M., Fan, H., Kitani, K.: Going deeper into first-person activity recognition. In: CVPR (2016)
23. McCandless, T., Grauman, K.: Object-centric spatio-temporal pyramids for egocentric activity recognition. In: BMVC (2013)
24. Miller, G.: Wordnet: a lexical database for english. Communications of the ACM (1995)
25. Moghimi, M., Azagra, P., Montesano, L., Murillo, A., Belongie, S.: Experiments on an rgb-d wearable vision system for egocentric activity recognition. In: CVPRW (2014)
26. Motwani, T., Mooney, R.: Improving video activity recognition using object recognition and text mining. In: ECAI (2012)
27. Ordonez, V., Liu, W., Deng, J., Choi, Y., Berg, A., Berg, T.: Predicting entry-level categories. IJCV (2015)
28. Pirsiavash, H., Ramanan, D.: Detecting activities of daily living in first-person camera views. In: CVPR (2012)
29. Ren, X., Gu, C.: Figure-ground segmentation improves handled object recognition in egocentric video. In: CVPR (2010)
30. Ryoo, M., Rothrock, B., Matthies, L.: Pooled motion features for first-person videos. In: CVPR (2015)
31. Sánchez, J., Perronnin, F., Mensink, T., Verbeek, J.: Image classification with the fisher vector: Theory and practice. IJCV (2013)
32. Sermanet, P., Eigen, D., Zhang, X., Mathieu, M., Fergus, R., LeCun, Y.: Overfeat: Integrated recognition, localization and detection using convolutional networks. ICLR (2013)

33. Singh, S., Arora, C., Jawahar, C.: First person action recognition using deep learned descriptors. In: CVPR (2016)
34. Spriggs, E., De La Torre, F., Hebert, M.: Temporal segmentation and activity classification from first-person sensing. In: CVPRW (2009)
35. Sundaram, S., Mayol-Cuevas, W.: Egocentric visual event classification with location-based priors. In: ISVC (2010)
36. Taralova, E., De La Torre, F., Hebert, M.: Source constrained clustering. In: ICCV (2011)
37. Wang, H., Kläser, A., Schmid, C., Liu, C.: Action recognition by dense trajectories. In: CVPR (2011)
38. Wang, H., Schmid, C.: Action recognition with improved trajectories. In: ICCV (2013)